The Throughput of Hybrid-ARQ in Block Fading under Modulation Constraints

Tarik Ghanim and Matthew C. Valenti
Lane Dept. of Comp. Sci. and Elect. Eng.
West Virginia University
Morgantown, WV 26506–6109
Email: mvalenti@wvu.edu

***Abstract*— In a seminal paper published in 2001, Caire and Tuninetti derived an information theoretic bound on the throughput of hybrid-ARQ in the presence of block fading. However, because the results placed no constraints on the modulation used, the input to the channel was Gaussian. The purpose of this paper is to investigate the impact of modulation constraints on the throughput of hybrid-ARQ in a block fading environment. First, we consider the impact of modulation constraints on information outage probability for a block fading channel with a fixed rate codeword. Then, we consider the effect of modulation constraints upon the throughput of hybrid-ARQ, where the rate of each codeword varies depending on the instantaneous channel conditions. These theoretical bounds are compared against the simulated performance of HSDPA, a newly standardized hybrid-ARQ protocol that uses QPSK and 16-QAM bit interleaved turbo-coded modulation. The results indicate how much of the difference between HSDPA and the previous unconstrained modulation bound is due to the use of the turbo-code and how much is due to the modulation constraints.**

I. Introduction

Hybrid-ARQ is a technique for combining forward error correction (FEC) coding with an automatic repeat request (ARQ) protocol [1]. A message is encoded by a low rate mother code and then partitioned into several blocks. Blocks are sent one by one until enough information is accumulated at the destination to correctly decode the message. Often, the channel is uncorrelated from one block to the next, in which case a block fading model may be assumed. A key performance metric for hybrid-ARQ is its throughput, which is the number of bits conveyed per unit time. In [2], Caire and Tuninetti derived information-theoretic bounds on the throughput of hybrid-ARQ in block fading. The results built upon related work on the performance of standard block fading channels [3], [4], i.e. channels with a fixed codeword size and number of blocks per codeword. The results in [2] placed no constraints upon modulation, and as a consequence, the input to the channel was assumed to be Gaussian distributed. However, practical systems do not use Gaussian-distributed modulation, and the computation of information-theoretic limits on the throughput of hybrid-ARQ *with practical modulation constraints* has until now remained an open problem.

The main motivation behind the present paper is the emergence of the High Speed Data Packet Access (HSDPA) standard [5], [6], which is part of the UMTS family of standards under development by the Third Generation Partnership Project (3GPP). In HSDPA, messages are first encoded with a binary turbo code and then punctured by a rate matching algorithm to create the first transmitted block. If the destination is unable to decode the initial block, then the codeword is again punctured by the rate matching algorithm, though by selecting a different set of rate matching parameters, a different set of code bits can be included in the second transmitted block. Blocks continue to be generated by rate matching with different parameters and sent until either the destination correctly decodes the message or an upper limit on the number of retransmissions is reached.

HSDPA uses either QPSK or (gray-labelled) 16-QAM modulation. Because the encoder is binary and separated from the modulator by a bitwise interleaver, this is an example of bit-interleaved coded-modulation (BICM) [7]. As shown in [7], the performance of a BICM-constrained system can differ significantly from that of a system with an unconstrained Gaussian input, especially at high spectral efficiency. The purpose of this paper is to investigate how modulation constraints effect performance of block fading channels in general (an issue that has also been recently discussed in [8]), and more specifically, the throughput of hybrid-ARQ over a block fading channel.

To accomplish this goal, we first begin in Section II with an exposition of our system model, and then continue in Section III with a review of the BICM-constrained capacity of simple additive white Gaussian noise (AWGN) channels. Section IV discusses the information outage probability of block fading with both unconstrained and constellation-constrained inputs, and Section V builds upon these results to derive the throughput and latency of hybrid-ARQ under modulation constraints, thereby generalizing the results of [2]. Numerical results in Section VI compare the simulated throughput of HSDPA against the unconstrained bound of [2] and the modulation-constrained bound developed in this paper. Finally conclusions and suggestions for future work are given in Section VII.

II. System Model

The system model is as shown in Fig. 1. The system uses bit-interleaved coded modulation [7] and hybrid-ARQ [2]. The transmitter passes a length K binary message $\mathbf{u}$ into a binary encoder, producing a codeword $\mathbf{c}'$ of length N bits. The codeword is bitwise interleaved, producing the vector $\mathbf{c}$, which

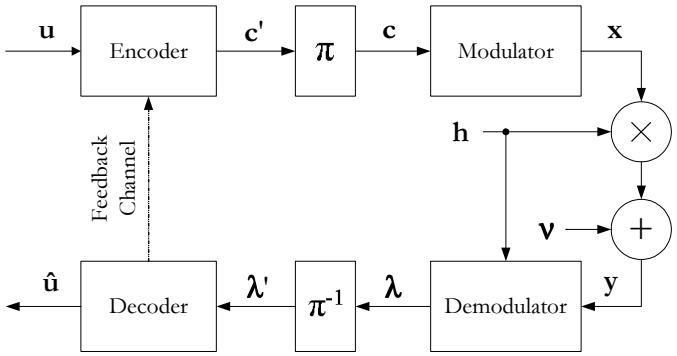

Fig. 1. System model. π denotes interleaving at the bit level.

is passed into an M-ary modulator. The modulator produces a length $\lceil N/\log_2 M \rceil$ vector $\mathbf{x}$ of complex M-ary symbols drawn from the signal set $\mathcal{S}$. The modulated codeword is broken into B_{max} equal-length blocks, denoted $\mathbf{x}[b], 1 \leq b \leq B_{max}$. The length of each block is $L = \lceil N/(B_{max} \log_2 M) \rceil$ symbols and the rate of each block is $R = K/L$.

The transmitter sends the first block $\mathbf{x}[1]$, and if the receiver is able to successfully decode it, an acknowledgement will be sent back through a feedback channel (we assume here that the feedback channel is error and delay-free and that an ideal error detecting code allows the receiver to discriminate between correctly and incorrectly decoded messages). If the transmitter receives an acknowledgement, it will move on to the next message; otherwise, it will send the next block from the current message. This process continues until either the message is received correctly or the last (B_{max}) block is transmitted.

The b^{th} block is transmitted with average energy per symbol $\mathcal{E}_s = E\{|x|^2\}$ over a block fading channel so that the received signal is:

$$\mathbf{y}[b] \quad = \quad h[b]\mathbf{x}[b] + \boldsymbol{\nu} \tag{1}$$

where $\boldsymbol{\nu}$ is a vector of complex Gaussian noise whose dimensions match $\mathbf{x}[b]$ and whose components are zero-mean i.i.d. circularly symmetric Gaussian with variance $N_o/2$ in each complex direction, and $h[b]$ is a complex scalar channel gain assumed to be independent from block to block and constant for the duration of each block. Without loss of generality, $E\{|h[b]|^2\} = 1$ so that the average received energy per symbol is the same as the transmitted symbol energy.

Each received symbol in $\mathbf{y}[b]$ is passed through a demodulator that produces log-likelihood ratio estimates of each of the $\log_2 M$ bits associated with the symbol. Since demodulation is on a symbol-by-symbol basis, consider the demodulation process for a single symbol y. For each possible $x_m, 1 \leq m \leq M$, a log-likelihood is formed:

$$\begin{aligned} \Lambda_m &= \log p(x_m|y) \\ &= \log \frac{p(x_m|y)}{\sum_{x\in\mathcal{S}} p(x|y)} \end{aligned} \tag{2}$$

where $p(x)$ is the pdf of x. Letting the likelihood $f(x|y) = \kappa p(x|y)$ for any arbitrary constant κ that is common for all postulated symbols, and applying Bayes' rule, then (2) can be more conveniently rewritten as

$$\begin{aligned} \Lambda_m &= \log \frac{f(y|x_m)}{\sum_{x\in\mathcal{S}} f(y|x)} \\ &= \log f(y|x_m) - \log \sum_{x\in\mathcal{S}} f(y|x) \\ &= \log f(y|x_m) - \max_{x\in\mathcal{S}}{}^{*} \left[\log f(y|x)\right] \end{aligned} \tag{3}$$

where the *max-star* operator is as defined in [9],

$$\max_i{}^{*} \{x_i\} \quad = \quad \log\left\{\sum_i e^{x_i}\right\}. \tag{4}$$

Coherent detection is implemented by using:

$$\log f(y|x) \quad = \quad -\frac{\mathcal{E}_s}{N_o}|y - hx|^2. \tag{5}$$

Notice in Fig. 1 that the receiver has perfect channel state information (CSI) but that the transmitter does not use any CSI (aside from the ACK signal sent over the feedback channel).

Next, the receiver transforms the set of M log-likelihoods that are calculated for each received symbol into a set of $\log_2 M$ bitwise log-likelihood ratios (LLRs), one for each code bit associated with the symbol. To calculate the LLR for the i^{th} bit of received symbol y, first partition the symbol set $\mathcal{S}$ into two disjoint sets, $\mathcal{S}_i^{(0)}$, which is the set of symbols whose i^{th} bit is a 0, and $\mathcal{S}_i^{(1)}$, which is the set of symbols whose i^{th} bit is a 1. The LLR of the i^{th} bit, $1 \leq i \leq \log_2 M$, is then:

$$\begin{aligned} \lambda_i &= \log \frac{p(c_i = 1|y)}{p(c_i = 0|y)} \\ &= \log \frac{\sum_{x\in\mathcal{S}_i^{(1)}} p(x|y)}{\sum_{x\in\mathcal{S}_i^{(0)}} p(x|y)}. \end{aligned} \tag{6}$$

When symbols are equally likely, this may be expressed as

$$\lambda_i \quad = \quad \max_{x\in\mathcal{S}_i^{(1)}}{}^{*} \left[\log f(y|x)\right] - \max_{x\in\mathcal{S}_i^{(0)}}{}^{*} \left[\log f(y|x)\right]. \tag{7}$$

After the b^{th} block has been received, then the corresponding bit likelihoods for all blocks that have been received so far are passed into a decoder. The blocks could be encoded in such a way that all B_{max} blocks are identical (a *repetition* code), in which case the blocks will be *diversity*-combined at the receiver by adding up the LLR's of each block. More generally, *incremental redundancy* could be used, whereby each block is obtained by puncturing a low rate mother code. With incremental redundancy, a different part of the codeword is transmitted each time, and after the b^{th} block, a receiver will pass the rate $R_b = R/b$ code that it has until then received through its decoder (*code-combining*).

III. AWGN Capacity

The mutual information between channel input X and output Y is defined as [10]:

$$I(X,Y) \quad = \quad \int\int p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} dxdy. \tag{8}$$

The capacity of a channel is found by maximizing the mutual information over all possible input distributions:

$$C = \max_{p(x)} I(X,Y). \quad (9)$$

When there are no constraints on the input signal and the channel is AWGN, (9) is maximized by letting the input $p(x)$ take on a Gaussian distribution. This results in the classic unconstrained AWGN channel capacity:

$$C(\gamma) = \log_2(1+\gamma) \quad (10)$$

where $\gamma = \mathcal{E}_s/N_o$ is the SNR and the capacity takes on units of bits per channel use (i.e. transmitted symbol).

Rather than using Gaussian distributed symbols, practical systems use symbols drawn from the signal set $\mathcal{S}$, usually with equal probability. Under such modulation constraints, $p(x)$ is a fixed function of $\mathcal{S}$, and since there is nothing to maximize over, the capacity is merely the mutual information given by (8) with $p(x)$ determined by the modulation constraint.

After some manipulation, (8) and (9), can be written in terms of the symbol likelihood Λ_m as the expectation

$$\begin{aligned} C(\gamma) &= E_{x_m,\nu}[\log M + \log p(x_m|y)] \\ &= \log M + E_{x_m,\nu}[\Lambda_m] \text{ nats/symbol} \\ &= \log_2 M + \frac{E_{x_m,\nu}[\Lambda_m]}{\log 2} \text{ bits/symbol} \end{aligned} \quad (11)$$

where the expectation is over all symbols $x_m \in \mathcal{S}$ and complex noise samples ν with SNR equal to γ. This expression represents the *coded modulation* (CM) capacity and can be evaluated either by numerical integration [3], [11] or Monte Carlo integration [7].

If the system is further constrained to use BICM [7], then the channel is essentially transformed into $\log_2 M$ parallel binary channels. The capacity of the i^{th} binary channel, $1 \le i \le \log_2 M$ is

$$C_i(\gamma) = E_{c_i,\nu}[\log 2 + \log p(c_i|y)] \text{ nats/symbol} \quad (12)$$

where the expectation is over the two possible code bits $c_i \in \{0,1\}$ and the complex noise samples ν with SNR γ. After some manipulation, this can be expressed in terms of the binary LLR λ_i as:

$$\begin{aligned} C_i(\gamma) &= \log(2) - E_{c_i,\nu}[\max*\{0,(-1)^{c_i}\lambda_i\}] \text{ nats/symbol} \\ &= 1 - \frac{E_{c_i,\nu}[\max*\{0,(-1)^{c_i}\lambda_i\}]}{\log 2} \text{ bits/symbol.} \end{aligned} \quad (13)$$

Since the capacities of parallel Gaussian channels add [10], the overall capacity of the BICM system is found by adding the capacities of the individual binary channels:

$$C(\gamma) = \sum_{i=1}^{\log_2 M} C_i(\gamma). \quad (14)$$

As an example, the capacity when $\mathcal{S}$ is constrained to be either QPSK or 16-QAM is shown in Fig. 2. For comparison purposes, the unconstrained capacity (10) is also shown. For 16-QAM, both the CM and BICM capacities are shown. While the CM capacity does not depend on how bits are mapped to symbols, for BICM it does. The BICM-constrained capacity for two typical symbol mappings are shown, gray-labelling and set-partitioning (SP). While both BICM capacities are inferior to the CM capacity, the BICM capacity with gray-labelling is very close to the CM capacity, especially for $C(\gamma) > 2$.

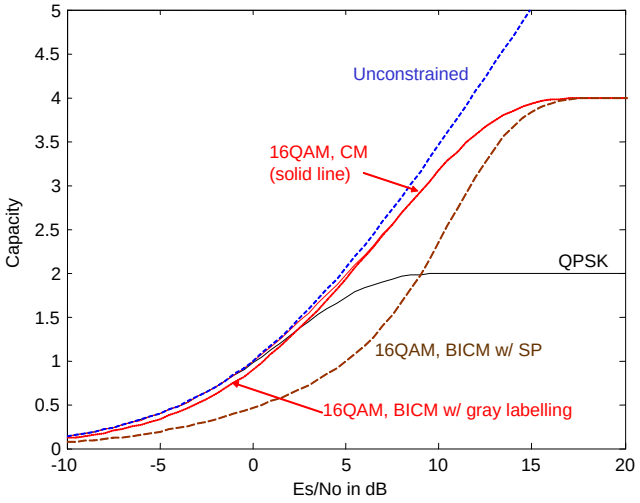

Fig. 2. Capacity of QPSK, 16-QAM, and unconstrained (Gaussian-input) modulation in AWGN. For 16-QAM, the CM capacity is shown as is the BICM capacities for two types of symbol mappings (gray-labelling and set-partitioning (SP)).

IV. Block Fading

In block fading, the codeword is broken into B blocks and each block is sent over an independent channel. Because the fading coefficient $h[b]$ of the b^{th} block is constant for the entire duration of the block, the channel during one block is conditionally Gaussian (conditioned on $h[b]$). However, since the fading coefficient is random, then so is the *instantaneous* SNR of the b^{th} block, which we denote $\gamma_b \equiv |h[b]|^2\mathcal{E}_s/N_0$, and therefore so is the corresponding capacity $C(\gamma_b)$. For Rayleigh block fading, $|h[b]|$ is Rayleigh and $|h[b]|^2$ is exponentially distributed. When code-combining is used, then the capacities of the B blocks add, since each block is sent over an independent Gaussian channel. The resulting capacity is:

$$C(\gamma_1,...,\gamma_B) = \frac{1}{B}\left(\sum_{b=1}^{B} C(\gamma_b)\right) \quad (15)$$

where the $\frac{1}{B}$ term is needed because blocks are orthogonal and therefore effectively occupy only $1/B^{th}$ of the channel.

For diversity combining, the SNRs add and so the capacity when B blocks are transmitted is:

$$C(\gamma_1,...,\gamma_B) = \frac{1}{B}C\left(\sum_{b=1}^{B}\gamma_b\right). \quad (16)$$

When there are no modulation constraints, the capacities in (15) and (16) are found from the unconstrained AWGN capacity (10), while when there are modulation constraints equation (11) or equations (13) and (14) must be used for CM and BICM, respectively.

When B is finite, the channel is not ergodic, and therefore a Shannon-sense channel capacity does not exist. For finite B, a more relevant performance metric is the *information outage probability*, defined in [3] and [4] as the probability that the instantaneous capacity $C(\gamma_1, ..., \gamma_B)$ is less than the rate $R_B = R/B$,

$$p_0(B) \quad = \quad P\left[C(\gamma_1, ..., \gamma_B) < R_B\right]. \tag{17}$$

Whenever $C(\gamma_1, ..., \gamma_B) < R_B$, an *information outage* occurs, and reliable signaling is not possible. The information outage probability is an information theoretic bound on the *frame error rate* (FER) in block fading, and thus no system can have a FER that is better than the information outage probability.

In the example shown in Fig. 3, the information outage probability of code-combining in Rayleigh block fading is plotted against SNR for rate $R_B = 2$ bits per symbol and $B = \{1, 2, 3, 4, 10\}$. For each value of B, two curves are shown, one for an unconstrained Gaussian input [obtained by substituting (10) into (15) with $R_B = 2$], and the other for a BICM constrained input using gray-labelled 16-QAM [obtained by substituting (13) into (14) and (15)]. On this log-log scale, each curve becomes a straight line at high SNR. The slope of the line is $-d$, where d is an integer in $[0, B]$ and is called the *block diversity* or *SNR exponent*. As discussed in [8], for an unconstrained Gaussian input channel, $d = B$, but under modulation constraints the diversity is upper-bounded by the Singleton bound

$$d \quad = \quad 1 + \left\lfloor B\left(1 - \frac{R_B}{\log_2 M}\right)\right\rfloor. \tag{18}$$

Since in this case $R_B/\log_2 M = 1/2$, $d = 1, 2, 2, 3$ and 6 for $B = 1, 2, 3, 4$ and 10, respectively. This behavior can be observed in the figure. For $B = 1$ and 2, the outage probability under modulation constraints is worse than the unconstrained case, but asymptotically the two curves for the same value of B have the same slope. However, for $B = 3$ not only is the constrained case worse than the unconstrained case, but asymptotically it has the same slope as the $B = 2$ unconstrained case. Similarly, the $B = 4$ constrained case has the same slope as the $B = 3$ unconstrained case. For $B = 10$, the asymptotic slope for the constrained case is indeed 6, though this is not obvious by looking at the figure because slope 6 and 10 look similar to the eye.

V. Hybrid-ARQ

Let the random variable B indicate the number of hybrid-ARQ transmissions until the packet is successfully received. Initially, consider the case that there is no limit on the number of transmissions. For B to equal b, the first $b-1$ attempts must fail while the b^{th} attempt must succeed. Thus, the pmf of B is

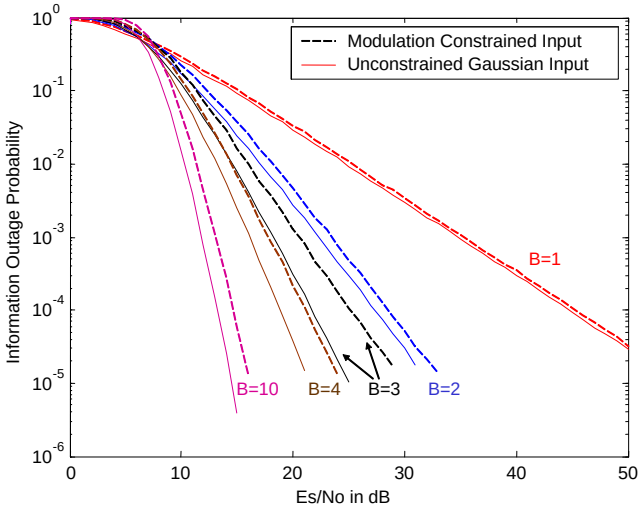

Fig. 3. Information outage probability vs. SNR for unconstrained and gray-labelled 16QAM modulation in Rayleigh block fading with rate $R_B = 2$.

$$p_B[b] \quad = \quad (1 - p_0(b)) \prod_{i=1}^{b-1} p_0(i) \quad \text{for } b \geq 1. \tag{19}$$

Often, an upper limit B_{max} is placed on the number of hybrid-ARQ transmissions. If the message is not received after B_{max} blocks have been transmitted, then an error is logged, and the system moves on to the next message. The pmf of B with constraint B_{max} on the number of transmissions is

$$p_B[b] \quad = \quad \begin{cases} \xi\,(1 - p_0(b)) \prod_{i=1}^{b-1} p_0(i) & \text{for } 1 \leq b \leq B_{max} \\ 0 & \text{otherwise,} \end{cases} \tag{20}$$

where ξ is a normalization factor required to make $p_B[b]$ a valid pmf:

$$\xi \quad = \quad \left[\sum_{i=1}^{B_{max}} (1 - p_0(b)) \prod_{i=1}^{b-1} p_0(i)\right]^{-1}. \tag{21}$$

Let τ be the time between the start of consecutive blocks (which includes the time to transmit the block, process it, send an acknowledgement, and process the acknowledgement). Then the throughput, in bits per second, is:

$$\eta \quad = \quad \frac{K}{\tau E[B]} \tag{22}$$

where $E[B]$ is the expected value of B, and K is the number of information bits per message. A more meaningful metric is the *throughput efficiency*, which is the ratio of *correct* bits to transmitted bits:

$$\eta_{eff} \quad = \quad \frac{1 - p_0(B_{max})}{E[B]}. \tag{23}$$

TABLE I

MAXIMUM THROUGHPUT (*kbps*) FOR THE FIXED REFERENCE CHANNEL

	QPSK	16-QAM
H-Set 1	534	777
H-Set 2	801	1166
H-Set 3	1601	2332

Another metric of interest is the latency, which is the time between correctly decoded messages, and is given by τ/η_{eff} seconds.

Note that when using hybrid-ARQ, $R_B = R/B$ and so the upper-bound on diversity given by (18) becomes

$$d \quad = \quad 1 + \left\lfloor B - \frac{R}{\log_2 M} \right\rfloor . \tag{24}$$

This implies that as long as $R < \log_2 M$ (which must be true in practical systems) then d is upper bounded by B and there is no loss in diversity in hybrid-ARQ systems due to using modulation constraints.

VI. LIMITS ON THE THROUGHPUT OF HSDPA

In this section, the throughput efficiency of HSDPA (obtained through computer simulations) is compared against the corresponding information theoretic bounds (both unconstrained and modulation-constrained). With HSDPA, the message is first encoded by the rate $1/3$ UMTS turbo code [12]. A two stage rate matching algorithm is used to puncture the codeword, which is then modulated (after bitwise interleaving) using either QPSK or 16-QAM. For each modulation type, there are eight ways to perform rate matching, which is specified by a three bit variable called the *redundancy version* [5]. In the case of 16-QAM, gray-labelling is used and rate matching can be used to essentially rearrange the signal constellation mapping. When a retransmission is requested, the rate matching algorithm can either be run with the same redundancy version, resulting in a repetition code which is diversity-combined at the receiver, or a different redundancy version can be used for each transmission, in which case code-combining is used.

Key parameters, such as the message size (K), block length after rate matching (L), sequence of redundancy versions, and time between the start of consecutive blocks (τ), were chosen to comply with the 3GPP approval standard [13]. There are a total of six testsets defined in [13], termed H-Set 1 through H-Set 6. In this section, we give throughput results for H-Set 1 through 3, which differ only in the value of τ. The maximum throughput for these three H-Sets, which occurs as $E[B] \to 1$, or equivalently as $\mathcal{E}_s/N_o \to \infty$, is given in Table I. For each case, the number of information bits is $K = 3202$ for QPSK and $K = 4664$ for 16-QAM. After rate matching, the block size is $L = 2400$ QPSK symbols or $L = 1920$ 16-QAM symbols. The maximum number of hybrid-ARQ transmissions per message is $B_{max} = 4$ and each block is punctured using a different redundancy version (code-combining). The time between the start of consecutive blocks is $\tau = 6$, 4, and 2 msec for H-Set 1, 2, and 3, respectively.

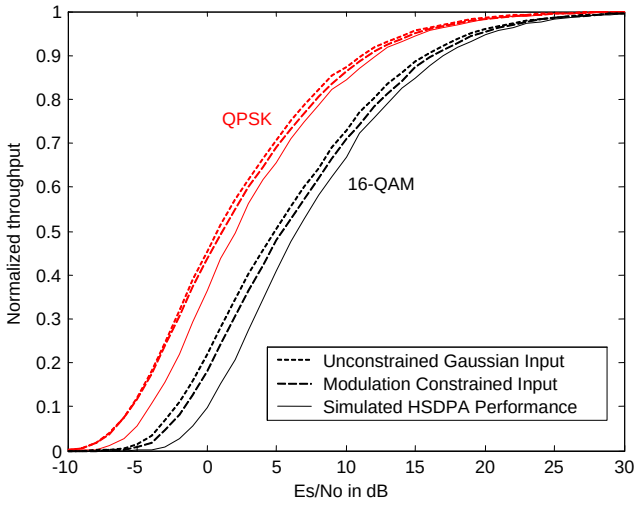

Fig. 4. Throughput efficiency of HSDPA H-Sets 1 through 3 in Rayleigh block fading using QPSK or 16-QAM modulation. For each modulation type, the unconstrained and modulation-constrained theoretical limits are compared against the simulated performance of the HSDPA system.

Fig. 4 shows throughput efficiency versus SNR in Rayleigh block fading for H-Sets 1 through 3 using both QPSK and 16-QAM modulation. Since H-Sets 1 through 3 differ only in the value of τ, all three have the same throughput efficiency. The figure shows two groups of three curves. The group on the left is for QPSK and the group on the right is for 16-QAM. Note that QPSK has better throughput efficiency than 16-QAM in this application because it has a lower per-block code rate ($R = 3202/2400$ for QPSK and $4664/1920$ for QAM). For each modulation type, three curves are shown. The leftmost curve is the information-theoretic limit on throughput with an unconstrained (i.e. Gaussian-distributed) input, while the middle curve is the information-theoretic limit with a modulation-constrained input. The rightmost curve is the throughput of the simulated HSDPA system in block fading.

The results shown in Fig. 4 indicate how much of the performance difference between HSDPA and the corresponding theoretical limits is due to the modulation constraints and how much is due to the use of the turbo code. For instance, with QPSK, a throughput efficiency $\eta_{eff} = 0.5$ is achieved at $\mathcal{E}_s/N_0 = 0.77$, 1.12, and 2.05 dB for the unconstrained, modulation-constrained, and actual HSDPA cases, respectively. This implies that while HSDPA has a 1.28 dB loss compared to the unconstrained theoretic bound, about 0.35 dB of this loss can be attributed to the modulation constraint, while the rest is attributed to the turbo code. Similarly, for 16-QAM, a throughput efficiency $\eta_{eff} = 0.5$ is achieved at $\mathcal{E}_s/N_o = 4.88$, 5.44, and 6.48 dB for the unconstrained, modulation-constrained, and actual HSDPA cases, respectively. This indicates that of the 1.60 dB difference between HSDPA and the unconstrained theoretic bound, about 0.56 dB of this loss is due to the modulation constraint. It is interesting to note that for QPSK, the loss due to the

modulation constraint diminishes at low throughput efficiency (e.g. $\eta_{eff} < 0.2$), while for QAM it does not (except at extremely small η_{eff}). These results suggest that the modulation constraint has more of a negative effect when using QAM signaling than when using QPSK signalling, at least for the HSDPA system considered here.

VII. Conclusions

When examining the throughput of any practical hybrid-ARQ system in block fading, there is always a loss relative to the information theoretic bounds derived by Caire and Tuninetti [2]. In the case of HSDPA, this loss is in the range of 1-2 dB. While there are several causes for this loss, these causes can be roughly partitioned into those that are due to the modulation constraints and those that are due to the use of a practical code. This paper presented a methodology for determining the information theoretic throughput bound under modulation constraints, thereby allowing the relative throughput losses due to modulation and coding to be separated. In the case of HSDPA, about 0.5 to 0.6 dB of the loss is due to using a 16-QAM modulation constraint, while up to 0.4 dB of the loss is due to using QPSK modulation constraints.

As for the losses due to causes other than modulation, there are several factors. First, both the unconstrained and modulation-constrained throughput bounds were found by using expressions for the AWGN Shannon-sense capacity of each block. As such, these expressions are derived under the assumption of an infinite block length. However, practical systems must use a finite block length (e.g. in HSDPA it is 2400 QPSK symbols or 1920 QAM symbols). Thus some of the loss is due to finite block length effects, and the amount of this loss can be determined using an extension of the sphere-packing approaching described in [14]. Another issue with HSDPA is that while the rate matching algorithm can be used to produce up to eight distinct blocks for each modulation type, these blocks are not mutually exclusive, i.e. some code bits will appear in more than one block. As a consequence, the processing at the receiver will actually be a combination of code-combining and diversity-combining. This problem can be alleviated by using a rate compatible code, such as a rate compatible turbo code [15], which will have distinct blocks and is therefore amenable to pure code-combining. One weakness of using rate compatible coding is that it imposes a finite upper limit on the maximum number of retransmissions B_{max}; this drawback can possibly be alleviated by using a rateless code such as an LT code [16] or a Raptor code [17]. In addition to finite block length effects and presence of repeated code bits, the other losses relative to the information theoretic bounds can be attributed to the *code imperfectness* as defined by [18].

While the purpose of this paper has been to examine the effects of modulation constraints upon the theoretical throughput limits of conventional, *point-to-point*, hybrid-ARQ, the results can easily be extended to study hybrid-ARQ based *relaying* protocols, such as the HARBINGER protocol proposed in [19]. In a relaying network, additional relay terminals assist the transmission of the message from source to destination. While the initial hybrid-ARQ transmission must always come from the source, each retransmissions may come from any relay that overhears the message. Thus the time-diversity benefits of hybrid-ARQ are combined with the spatial-diversity of relaying. While the results in [19] assumed an unconstrained channel input, the results from this paper could be used to study the impact of modulation constraints on hybrid-ARQ relaying protocols.

Acknowledgment

The authors would like to thank Dr. Mike McCloud from Tensorcomm, Inc., and Shi Cheng and Rohit Iyer Seshadri from WVU's Wireless Communications Research Lab for their technical guidance.

References

[1] S. Wicker, *Error Control Systems for Digital Communications and Storage*. Englewood Cliffs, NJ: Prentice Hall, Inc., 1995.

[2] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Trans. Inform. Theory*, vol. 47, no. 5, pp. 1971–1988, July 2001.

[3] R. Knopp and P. A. Humblet, "On coding for block fading channels," *IEEE Trans. Inform. Theory*, vol. 46, no. 1, pp. 189–205, Jan. 2000.

[4] L. Ozarow, S. Shamai, and A. D. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. Veh. Tech.*, vol. 43, pp. 359–378, May 1994.

[5] European Telecommunications Standards Institute, "Universal mobile telecommunications system (UMTS): Multiplexing and channel coding (FDD)," *3GPP TS 125.212 version 6.6.0, release 6*, Sept. 2005.

[6] R. Love, A. Ghosh, W. Xiao, and R. Ratasuk, "Performance of 3GPP high speed downlink packet access (HSDPA)," in *Proc. IEEE Veh. Tech. Conf. (VTC)*, Los Angeles, Sept. 2004.

[7] G. Caire, G. Taricco, and E. Biglieri, "Bit-interleaved coded modulation," *IEEE Trans. Inform. Theory*, vol. 44, pp. 927–946, May 1998.

[8] A. G. Fàbregas and G. Caire, "Coded modulation in the block-fading channel: Coding theorems and code construction," *IEEE Trans. Inform. Theory*, vol. 52, pp. 91–114, Jan. 2006.

[9] A. J. Viterbi, "An intuitive justification and a simplified implemetation of the MAP decoder for convolutional codes," *IEEE J. Select. Areas Commun.*, vol. 16, no. 2, pp. 260–264, Feb. 1998.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 1991.

[11] S. G. Wilson, *Digital Modulation and Coding*. Upper Saddle River, NJ: Prentice Hall, 1996.

[12] M. C. Valenti and J. Sun, "The UMTS turbo code and an efficient decoder implementation suitable for software defined radios," *Int. J. Wireless Info. Networks*, vol. 8, pp. 203–216, Oct. 2001.

[13] European Telecommunications Standards Institute, "Universal mobile telecommunications system (UMTS): User equipment (UE) radio transmission and reception (FDD)," *3GPP TS 125.101 version 6.9.0, release 6*, Sept. 2005.

[14] S. Dolinar, D. Divsalar, and F. Pollara, "Code performance as a function of block size," JPL TDA Progress Report, Tech. Rep., May 1998.

[15] D. N. Rowitch and L. B. Milstein, "On the performance of hybrid FEC/ARQ systems using rate compatable punctured turbo (RCPT) codes," *IEEE Trans. Commun.*, vol. 48, no. 6, pp. 948–959, June 2000.

[16] M. Luby, "LT codes," in *IEEE Symposium on Foundations of Computer Science*, Vancouver, Nov. 2002.

[17] A. Shokrollahi, "Raptor codes," in *Proc. IEEE Int. Symp. on Inform. Theory (ISIT)*, Chicago, July. 2004.

[18] S. Dolinar, D. Divsalar, and F. Pollara, "Turbo code performance as a function of code block size," in *Proc. IEEE Int. Symp. on Inform. Theory (ISIT)*, Boston, Aug. 1998.

[19] B. Zhao and M. C. Valenti, "Practical relay networks: A generalization of hybrid-ARQ," *IEEE J. Select. Areas Commun.*, vol. 23, no. 1, Jan. 2005.